Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1351 – 1360

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Canonical correlation analysis for geographical and chronological responses.

Mariko Yamamura[1,*], Hirokazu Yanagihara[1], Hiroko Kato Solvang[1], Nils Øien[1], Tore Haug[1]

[a] *Graduate School of Education, Hiroshima University, 1-1-1 Kagamiyama, Higashi-Hiroshima, 739-8524, Japan*
[b] *Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, 739-8626, Japan*
[c] *Institute of Marine Research, PO Box 1870 Nordnes, N-5817, Bergen, Norway*
[d] *Institute of Marine Research, PO Box 6404 N-9294, Tromsø, Norway*

Abstract

Data containing information about observed location and time are called geographical and chronological data. The purpose of this paper is to propose how we can analyze geographical and chronological data with multiple response variables by innovating the varying coefficient model in canonical correlation analysis. In addition, the variable selection proposed by Hashiyama *et al*. (2014) is applied to our model. As numerical background, we propose to apply an approach where we use a body condition data set from common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the Barents Sea (Solvang *et al*. (2016)). From the estimation results, minke whale body condition is affected by geography in females and by chronology in males, however the geographical effect seems not so strong, and male and female whales gain their body condition as fall approaches, which is the well known as their general habits in the Barents Sea.

Peer-review under responsibility of KES International

Keywords: Canonical correlation analysis; Multidimensional time-space data; Variable selection

1. Introduction

Data containing information about observed location and time are called geographical and chronological data. Yamamura *et al*. (2016) introduced application of a varying coefficient model to geographical and chronological data with one response variable, using the Japanese cedar (*Cryptomeria japonica*) data from Yoshimoto *et al*. (2012) which included longitudes / latitudes where the tree was planted as geographical variables and the age of the tree as a chronological variable. The varying coefficient model was originally proposed by Hastie and Tibshirani (1993), and Tonda *et al*. (2010) applied it for geographical data. Yamamura *et al*. (2016) extended the application potency of the varying coefficient model in Tonda *et al*. (2010) by applying the model not only for geographical, but also for

* Corresponding author. Tel.: +81-(0)82-424-4608 ; fax: +81-(0)82-424-3463.
E-mail address: yamamura@hiroshima-u.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.180

chronological data, and in practice showed how the difference in tree growth depended on the geographical location and the age of the tree.

As is often the case with real data sets, we sometimes need to analyze with multiple response variables. Tonda *et al*. (2010) has only proposed the varying coefficient model for a single response variable. One method of treating multiple response variables is that we create a synthesis variable from them and apply a regression model which procedure corresponds to canonical correlation analysis (CCA). The purpose of this paper is to propose how we can analyze geographical and chronological data with multiple response variables by innovating the varying coefficient model in CCA. In addition, the variable selection proposed by Hashiyama *et al*. (2014) is applied to our model since selecting effective variables from data is absolutely imperative for estimation. Leurgans *et al*. (1993) and Dubin and Müller (2005) proposed CCA for longitudinal data by functional approach and their models are more flexible than ours. Our model uses the basic CCA in Hotelling (1936), and its features are the estimation method is simple and the model is not just for longitudinal data. Then we can fit any kind of variables instead of time in our model, in fact, time and location variables are used in this paper. In addition more than one kind time variable could be used in our model, such as two time variables; month and year.

In CCA, we are interested in investigating relationships between two sets of variables $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_q)'$, where the notation " $'$ " means the transpose. The goal of CCA, as developed by Hotelling (1936), is to construct two new sets of canonical variates $u = \boldsymbol{\alpha}'\boldsymbol{y}$ and $v = \boldsymbol{\theta}'\boldsymbol{x}$ that are linear combinations of the original variables such that the simple correlation between u and v is maximal, subject to the restriction that each canonical variate u and v has unit variance and is uncorrelated with other constructed variates within the set. For more details about CCA, see e.g. Timm (2002). In our model, $y_1, \ldots, y_p$ are multiple response variables synthesized as u by the linear model, while $x_1, \ldots, x_q$ are variables related to geographical and chronological variables synthesized as v by the varying coefficient model, and parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are estimated to be the correlation between u and v is maximal as well as the original CCA.

As numerical background, we propose to apply an approach where we use a body condition data set from common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the Barents Sea (Solvang *et al*. (2016)). Minke whales are one of the most abundant cetacean species during summer in the Northeast Atlantic. Their migration pattern brings them from overwintering locations at lower latitudes where they are supposed to spend the energy deposited at high productive arctic latitudes in summer. It is therefore expected that their body condition on the summer grounds will reflect food availability during their most intensive feeding period and thus indicate how well the Barents Sea ecosystem can support the population. During the commercial catch operations in Norwegian waters, data have been collected from all animals caught from 1993 to 2013. The data collected include year, month (April to September), day, latitude / longitude, sex, girth, length and three blubber thickness measurements in millimeter (see Solvang *et al*. (2016)) We use the blubber thickness measured at three specific sites, the girth, and the length as describing the body condition.

In our proposed model, three body condition variables out of five available are used for multiple response variables $\boldsymbol{y}$. For geographical and chronological variables we used latitude / longitude as geographical information and year and calendar day as chronological information. Varying coefficients for the latitude / longitude are represented by a cubic plane curve, while those for year and calendar day are plane curves, and the differences in body condition depending on geographical point in the Barents Sea, year, and calendar day are assessed from the estimation result. In addition to the actual use of the proposed model, the results also provides a more complete picture of whale's nutrition conditions in Barents Sea in relation to year, and season from April to September.

This paper is organized as follows: in section 2, the model is presented, the application of the model to the minke whale data is presented in section 3, while section 4 contains our conclusion.

2. Estimation Method

Let $\boldsymbol{y} = (y_1, \ldots, y_p)'$ be a p-variate vector of response variables, and $\boldsymbol{a} = (a_1, \ldots, a_k)'$ be a k-variate vector of explanatory variables. When a regression of each component of $\boldsymbol{y}$ on $\boldsymbol{a}$ is considered, p-regression-equations are appeared and sometimes difficult to be interpreted simultaneously. Hence, in order to reduce p-regression-equations to the one, a regression of $\boldsymbol{\alpha}'\boldsymbol{y}$ can be considered, where $\boldsymbol{\alpha}$ is a p-variate vector of unknown parameters. In particular, we assume that regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ changes by $\boldsymbol{z}$, where $\boldsymbol{z}$ is an m-variate vector of geographical

and chronological variables, time and location (latitude / longitude), i.e., $\boldsymbol{\beta}$ is expressed as a functional coefficient with respect to $\boldsymbol{z}$. Consequently, the regression model is given by

$$\boldsymbol{\alpha}'\boldsymbol{y} = \mu + \boldsymbol{\beta}(\boldsymbol{z})'\boldsymbol{a} + \varepsilon, \quad E[\varepsilon] = 0 \text{ and } Var[\varepsilon] = \sigma^2, \tag{1}$$

where $\boldsymbol{\beta}(\boldsymbol{z}) = (\beta_1(\boldsymbol{z}), \ldots, \beta_k(\boldsymbol{z}))'$ and $\beta_1(\boldsymbol{z}), \ldots, \beta_k(\boldsymbol{z})$ are called varying coefficients. Here, we use r-functions $w_1(\boldsymbol{z}), \ldots, w_r(\boldsymbol{z})$ to represent a variability of $\boldsymbol{\beta}(\boldsymbol{z})$. For examples, we use a cubic polynomial function (see e.g. Tonda *et al.* (2010)) to represent the variation when $m = 1$, then $(w_1(z), w_2(z), w_3(z), w_4(z))' = (1, z, z^2, z^3)'$, and a quadratic polynomial function when $m = 2$, then $\boldsymbol{z} = (z_1, z_2)'$ and $(w_1(\boldsymbol{z}), w_2(\boldsymbol{z}), w_3(\boldsymbol{z}), w_4(\boldsymbol{z}))' = (1, z_1, z_2, z_1^2, z_2^2)'$. Let $\boldsymbol{w}(\boldsymbol{z}) = (w_1(\boldsymbol{z}), \ldots, w_r(\boldsymbol{z}))'$. By using r-variate vector of unknown parameters $\boldsymbol{\theta}_j$, the varying coefficient $\beta_j(\boldsymbol{z})$ is given as

$$\beta_j(\boldsymbol{z}) = \boldsymbol{w}(\boldsymbol{z})'\boldsymbol{\theta}_j, \quad (j = 1, \ldots, k).$$

Let $\boldsymbol{x}(\boldsymbol{z}) = \boldsymbol{a} \otimes \boldsymbol{w}(\boldsymbol{z})$, where the notation "$\otimes$" means the Kronecker product. Then, the $\boldsymbol{a} \otimes \boldsymbol{w}(\boldsymbol{z})$ is a kr-variate vector stacking $a_1\boldsymbol{w}(\boldsymbol{z})$ to $a_k\boldsymbol{w}(\boldsymbol{z})$, i.e., $\boldsymbol{x}(\boldsymbol{z}) = (a_1\boldsymbol{w}(\boldsymbol{z})', \ldots, a_k\boldsymbol{w}(\boldsymbol{z})')'$. Henceforth, we represent $kr = q$. It follows from an elementary linear algebra that

$$\boldsymbol{\beta}(\boldsymbol{z})'\boldsymbol{a} = \boldsymbol{w}(\boldsymbol{z})'(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)\boldsymbol{a} = \sum_{j=1}^{k} a_j\boldsymbol{w}(\boldsymbol{z})'\boldsymbol{\theta}_j = (\boldsymbol{\theta}_1' \ldots, \boldsymbol{\theta}_k')\begin{pmatrix} a_1\boldsymbol{w}(\boldsymbol{z}) \\ \vdots \\ a_k\boldsymbol{w}(\boldsymbol{z}) \end{pmatrix} = (\boldsymbol{\theta}_1' \ldots, \boldsymbol{\theta}_k')(\boldsymbol{a} \otimes \boldsymbol{w}(\boldsymbol{z})) = \boldsymbol{\theta}'\boldsymbol{x}(\boldsymbol{z}),$$

where $\boldsymbol{\theta}$ is a kr-variate vector of unknown parameters stacking $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_k$, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_k')'$. Hence, the model in (**??**) can be rewritten as a regression of $\boldsymbol{\alpha}'\boldsymbol{y}$ on $\boldsymbol{x}(\boldsymbol{z})$, i.e.,

$$\boldsymbol{\alpha}'\boldsymbol{y} = \mu + \boldsymbol{\theta}'\boldsymbol{x}(\boldsymbol{z}) + \varepsilon, \quad E[\varepsilon] = 0 \text{ and } Var[\varepsilon] = \sigma^2, \tag{2}$$

In (**??**), we assume that $\boldsymbol{x}(\boldsymbol{z})$ does not includes the constant 1.

Let $\{(\boldsymbol{y}_i, \boldsymbol{a}_i, \boldsymbol{z}_i) | \ i = 1, \ldots, n\}$ be n-observable-data-pairs on $\boldsymbol{y}$, $\boldsymbol{a}$ and $\boldsymbol{z}$. Then, we estimate $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and μ by minimizing the following residual sum of squares (RSS):

$$\mathrm{RSS}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mu) = \sum_{i=1}^{n} \left\{\boldsymbol{\alpha}'\boldsymbol{y}_i - \mu - \boldsymbol{\theta}'\boldsymbol{x}_i(\boldsymbol{z}_i)\right\}^2, \tag{3}$$

where $\boldsymbol{x}_i(\boldsymbol{z}_i) = \boldsymbol{a}_i \otimes \boldsymbol{w}(\boldsymbol{z}_i)$. Let $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{x}}$ be sample means of $\boldsymbol{y}$ and $\boldsymbol{x}(\boldsymbol{z})$, respectively, i.e., $\bar{\boldsymbol{y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i$ and $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{z}_i)$, and $\hat{\mu} = \boldsymbol{\alpha}'\bar{\boldsymbol{y}} - \boldsymbol{\theta}'\bar{\boldsymbol{x}}$. From a method of the single linear regression, we can see that $\hat{\mu}$ minimizes the RSS in (**??**) given $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. This implies that

$$\mathrm{RSS}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mu) \geq \mathrm{RSS}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \hat{\mu}) = \sum_{i=1}^{n} \left\{\boldsymbol{\alpha}'(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) - \boldsymbol{\theta}'(\boldsymbol{x}_i(\boldsymbol{z}_i) - \bar{\boldsymbol{x}})\right\}^2 = F(\boldsymbol{\alpha}, \boldsymbol{\theta}). \tag{4}$$

In order to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ in (**??**), we minimize the $F(\boldsymbol{\alpha}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. Let $\boldsymbol{S}_{yy}$ and $\boldsymbol{S}_{xx}$ be $p \times p$ and $q \times q$ variance-covariance matrices of $\boldsymbol{y}$ and $\boldsymbol{x}(\boldsymbol{z})$, respectively, and $\boldsymbol{S}_{yx}$ be a $p \times q$ covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{x}(\boldsymbol{z})$, i.e.,

$$\boldsymbol{S}_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})', \ \boldsymbol{S}_{yx} = \frac{1}{n-1}\sum_{i=1}^{n}\{\boldsymbol{y}_i - \bar{\boldsymbol{y}}\}\{\boldsymbol{x}_i(\boldsymbol{z}_i) - \bar{\boldsymbol{x}}\}', \ \boldsymbol{S}_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}\{\boldsymbol{x}_i(\boldsymbol{z}_i) - \bar{\boldsymbol{x}}\}\{\boldsymbol{x}_i(\boldsymbol{z}_i) - \bar{\boldsymbol{x}}\}'. \tag{5}$$

In order to ensure a possibility of estimating the model in (**??**), we assume that $\boldsymbol{\alpha}'\boldsymbol{S}_{yy}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\theta}'\boldsymbol{S}_{xx}\boldsymbol{\theta} = 1$. Hence, estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are given by

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\alpha}\in\mathcal{A},\ \boldsymbol{\theta}\in\mathcal{T}} F(\boldsymbol{\alpha}, \boldsymbol{\theta}), \quad \mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^p | \ \boldsymbol{\alpha}'\boldsymbol{S}_{yy}\boldsymbol{\alpha} = 1\}, \ \mathcal{T} = \{\boldsymbol{\theta} \in \mathbb{R}^q | \ \boldsymbol{\theta}'\boldsymbol{S}_{xx}\boldsymbol{\theta} = 1\}. \tag{6}$$

Then, estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ correspond to estimated coefficients in canonical correlation analysis of $\boldsymbol{y}$ and $\boldsymbol{x}(\boldsymbol{z})$, because solutions of the minimization problem in (**??**) is equivalent to those in the maximization problem in CCA (see **??**). Hence, the solutions $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ is given by

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{S}_{yy}^{-1/2}\boldsymbol{g}, \quad \hat{\boldsymbol{\theta}} = \boldsymbol{S}_{xx}^{-1/2}\boldsymbol{h}, \tag{7}$$

where $\boldsymbol{g}$ is an eigen vector corresponding to the maximum eigen value of $\boldsymbol{S}_{yy}^{-1/2}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{yx}'\boldsymbol{S}_{yy}^{-1/2}$ and $\boldsymbol{h}$ is an eigen vector corresponding to the maximum eigen value of $\boldsymbol{S}_{xx}^{-1/2}\boldsymbol{S}_{yx}'\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1/2}$. By using $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_k')'$, estimated varying coefficient is given by $\hat{\beta}_j(z) = \hat{\boldsymbol{\theta}}_j'\boldsymbol{w}(z)$ $(j = 1, \ldots, k)$.

To obtain optimal $\boldsymbol{\alpha}$ and varying coefficients $\boldsymbol{\beta}(z)$, we should apply variable selection for $\boldsymbol{y}$ and $\boldsymbol{a}$, and find the best fitted $\boldsymbol{w}(z)$. For this, we apply an approach for the selection of a redundancy model in CCA (see e.g., Fujikoshi *et al.*, 2008) as follows: Without loss of generality, we divide $\boldsymbol{y}$ and $\boldsymbol{x}$ into two sub-vectors $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2')'$ and $\boldsymbol{x}(z) = (\boldsymbol{x}_3', \boldsymbol{x}_4')'$, where $\boldsymbol{y}_1$ and $\boldsymbol{x}_3$ are p_1- and q_1-variate vectors of variables assumed to be required, respectively. Corresponding to the divisions of $\boldsymbol{y}$ and $\boldsymbol{x}(z)$, we divide $\boldsymbol{S}_{yy}$, $\boldsymbol{S}_{xx}$ and $\boldsymbol{S}_{yx}$ as

$$\boldsymbol{S}_{yy} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}' & \boldsymbol{S}_{22} \end{pmatrix}, \quad \boldsymbol{S}_{yx} = \begin{pmatrix} \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \\ \boldsymbol{S}_{23} & \boldsymbol{S}_{24} \end{pmatrix}, \quad \boldsymbol{S}_{xx} = \begin{pmatrix} \boldsymbol{S}_{33} & \boldsymbol{S}_{34} \\ \boldsymbol{S}_{34}' & \boldsymbol{S}_{44} \end{pmatrix}.$$

Then, we consider the following hypotheses:

H_1 : $\boldsymbol{y}_2$ and $\boldsymbol{x}_4$ are irrelevant, H_2 : $\boldsymbol{y}_2$ is irrelevant,
H_3 : $\boldsymbol{x}_4$ is irrelevant, H_4 : $\boldsymbol{y}_2$ and $\boldsymbol{x}_4$ are not irrelevant.

Then, BICs for assessing H_j $(j = 1, 2, 3, 4)$ are defined by

$$\mathrm{BIC} = \begin{cases} -(n-1)\log\dfrac{|\boldsymbol{S}_{(24)(24)\cdot(13)}|}{|\boldsymbol{S}_{22\cdot1}||\boldsymbol{S}_{44\cdot3}|} + p_1q_1\log n, & (\text{for } H_1) \\ -(n-1)\log\dfrac{|\boldsymbol{S}_{(2x)(2x)\cdot1}|}{|\boldsymbol{S}_{22\cdot1}||\boldsymbol{S}_{xx\cdot1}|} + p_1q\log n, & (\text{for } H_2) \\ -(n-1)\log\dfrac{|\boldsymbol{S}_{(y4)(y4)\cdot3}|}{|\boldsymbol{S}_{yy\cdot3}||\boldsymbol{S}_{44\cdot3}|} + pq_1\log n, & (\text{for } H_3) \\ pq\log n, & (\text{for } H_4) \end{cases},$$

where the notation "$|\boldsymbol{S}|$" means the determinant of a square matrix $\boldsymbol{S}$, and

$$\boldsymbol{S}_{22\cdot1} = \boldsymbol{S}_{22} - \boldsymbol{S}_{12}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}, \quad \boldsymbol{S}_{44\cdot3} = \boldsymbol{S}_{44} - \boldsymbol{S}_{34}'\boldsymbol{S}_{33}^{-1}\boldsymbol{S}_{34},$$

$$\boldsymbol{S}_{(24)(24)\cdot(13)} = \begin{pmatrix} \boldsymbol{S}_{22} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{24}' & \boldsymbol{S}_{44} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_{12} & \boldsymbol{S}_{14} \\ \boldsymbol{S}_{23}' & \boldsymbol{S}_{34} \end{pmatrix}' \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{13} \\ \boldsymbol{S}_{13}' & \boldsymbol{S}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{S}_{12} & \boldsymbol{S}_{14} \\ \boldsymbol{S}_{23}' & \boldsymbol{S}_{34} \end{pmatrix},$$

$$\boldsymbol{S}_{xx\cdot1} = \boldsymbol{S}_{xx} - \begin{pmatrix} \boldsymbol{S}_{13}' \\ \boldsymbol{S}_{14}' \end{pmatrix} \boldsymbol{S}_{11}^{-1} \begin{pmatrix} \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \end{pmatrix}, \quad \boldsymbol{S}_{(2x)(2x)\cdot1} = \begin{pmatrix} \boldsymbol{S}_{22} & \boldsymbol{S}_{23} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{23}' & \boldsymbol{S}_{33} & \boldsymbol{S}_{34} \\ \boldsymbol{S}_{24}' & \boldsymbol{S}_{34}' & \boldsymbol{S}_{44} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_{12}' \\ \boldsymbol{S}_{13}' \\ \boldsymbol{S}_{14}' \end{pmatrix} \boldsymbol{S}_{11}^{-1} \begin{pmatrix} \boldsymbol{S}_{12} & \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \end{pmatrix},$$

$$\boldsymbol{S}_{yy\cdot3} = \boldsymbol{S}_{yy} - \begin{pmatrix} \boldsymbol{S}_{13} \\ \boldsymbol{S}_{23} \end{pmatrix} \boldsymbol{S}_{33}^{-1} \begin{pmatrix} \boldsymbol{S}_{13}' & \boldsymbol{S}_{23}' \end{pmatrix}, \quad \boldsymbol{S}_{(y4)(y4)\cdot3} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \boldsymbol{S}_{14} \\ \boldsymbol{S}_{12}' & \boldsymbol{S}_{22} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{14}' & \boldsymbol{S}_{24}' & \boldsymbol{S}_{44} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_{13} \\ \boldsymbol{S}_{23} \\ \boldsymbol{S}_{34}' \end{pmatrix} \boldsymbol{S}_{33}^{-1} \begin{pmatrix} \boldsymbol{S}_{13}' & \boldsymbol{S}_{23}' & \boldsymbol{S}_{34} \end{pmatrix}.$$

3. Real Data Analysis

3.1. Data

Over the period 1993-2013, the body condition data were obtained from a total of 10,556 common minke whales taken in Norwegian scientific and commercial whaling operations in the Northeast Atlantic during the months April to September. Immediately after death, the whales were taken onboard and hauled across the fore-deck of the boat. Total body length was measured in a straight line from the tip of the upper jaw to the apex of the tail fluke notch; girth was measured right behind the flipper; and blubber thickness was measured at three sites

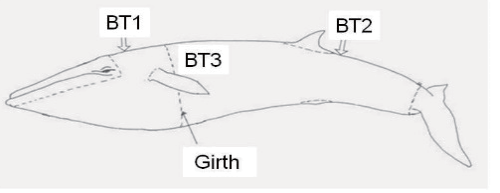

Fig. 1. Measurement sites.

(Fig.1): Dorsally behind the blowhole (BT1) and behind the dorsal fin (BT2), and laterally just above the center of the flipper (BT3). Blubber measurements were made perpendicular from the skin surface to the muscle-connective tissue interface. Length and girth measurements were made to the nearest centimeter, while blubber measurements were to the nearest millimeter. For all whales, the year, month, day, and latitude / longitude were recorded.

After removing data in period 1993-1996 and data with missing values, final numbers of individuals included in the analysis are 9,112 where 2,767 are males and 6,345 are females. We use the part of the data set from Solvang *et al*. (2016) (1997-2013), because the data from the years 1993-1996 are omitted due to uncertainties. Only the length, BT1, and BT3 are used as multiple response variables $\boldsymbol{y}$, whereas girth and BT2 are omitted from further analyses since they were reported to have potential measurement errors (Solvang *et al*. (2016)).

The data are described in Table 1. By comparing mean, minimum, and maximum values of length, BT1, and BT3, the body condition is not substantially different between the sexes, but it appear that females are slightly larger in size than males. From mean of latitude, 68.111 in male and 71.914 in female, females migrates further to the north than males. The calendar day counts the number of days from January 1 to the recorded date in the year, in order to see the seasonal effect in the estimation. The mean value 172.762 for males indicates that male whales appeared in the second half of May on average.

Table 1. Data description.

		Length (*cm*)	BT1 (*mm*)	BT3 (*mm*)	Latitude	Longitude	Calendar day (Jan.1=1)
Male	Mean	736.106	34.966	32.053	68.111	16.515	172.762
(2,767 obs.)	(S.D.)	93.756	8.990	7.760	4.894	9.136	25.320
	Min.	400.000	10.000	10.000	56.650	- 9.083	101.000
	Max.	925.000	90.000	80.000	79.550	35.033	259.000
Female	Mean	744.611	39.591	35.936	71.914	16.744	159.552
(6,345 obs.)	(S.D.)	98.535	10.308	9.239	5.824	9.721	20.801
	Min.	350.000	10.000	3.000	56.500	-9.133	100.000
	Max.	980.000	100.000	98.000	81.300	35.033	263.000

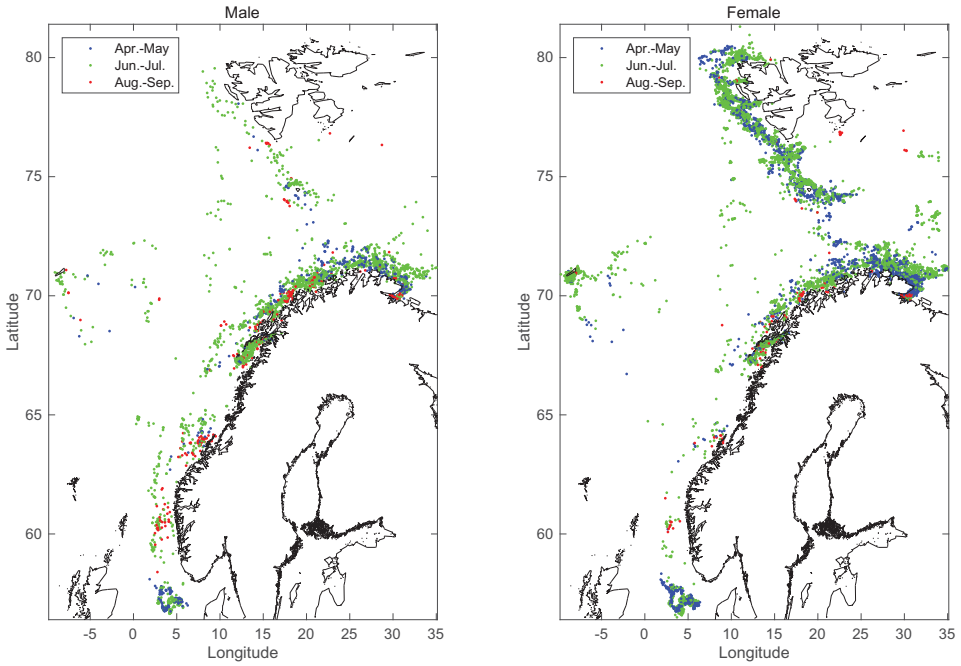

Fig. 2. Observations stratified by two-month periods.

Fig.2 shows the geographical positions where minke whales were caught, divided by sex, and the marker colors indicate the periods; "blue", "green", and "red" express "April-May", "June-July", and "August-September", respectively. Male and female whales have different migration preference areas such that more females are in the northmost areas around Svalbard (the island archipelago north of the Barents Sea) than males, which seldom appear as far north.

3.2. Results and Discussion

The synthesis variable $u = \boldsymbol{\alpha}'\boldsymbol{y}$ is assumed to have a liner structure, where $\boldsymbol{\alpha}$ is a parameter vector and columns of $\boldsymbol{y}$ are $(y_1, y_2, y_3)'$ = ("length", "BT1", "BT3")$'$. Explanatory variables $\boldsymbol{a}$ take values 1. We fit the linear model to estimate the varying coefficient cubic plane or plane curve, i.e. $\hat{\boldsymbol{\beta}}(z_1, z_2, z_3, z_4) = \hat{\boldsymbol{\theta}}'\boldsymbol{w}(z_1, z_2, z_3, z_4)$, where $(z_1, z_2, z_3, z_4)'$ = ("latitude", "longitude", "year", "calendar day")$'$. The $\boldsymbol{w}(z_1, z_2, z_3, z_4) = (\boldsymbol{w}_1(z_1, z_2)', \boldsymbol{w}_2(z_3)', \boldsymbol{w}_3(z_4))'$ is assumed to have one of either linear, quadratic or cubic expression with their interaction each, such as

$$\boldsymbol{w}_1(z_1, z_2) = \begin{cases} (z_1, z_2)' & (r_1 = 1) \\ (z_1, z_2,\ z_1^2, z_2^2, z_1 z_2)' & (r_1 = 2) \\ (z_1, z_2,\ z_1^2, z_2^2, z_1 z_2,\ z_1^3, z_2^3, z_1^2 z_2, z_1 z_2^2)' & (r_1 = 3) \end{cases}, \tag{8}$$

for "logitude" and "latitude", and those for "year" and "calendar day",

$$\boldsymbol{w}_2(z_3) = \begin{cases} (z_3)' & (r_2 = 1) \\ (z_3, z_3^2)' & (r_2 = 2) \\ (z_3, z_3^2, z_3^3)' & (r_2 = 3) \end{cases}, \tag{9}$$

and

$$\boldsymbol{w}_3(z_4) = \begin{cases} (z_4)' & (r_3 = 1) \\ (z_4, z_4^2)' & (r_3 = 2) \\ (z_4, z_4^2, z_4^3)' & (r_3 = 3) \end{cases}, \tag{10}$$

respectively, where r_d $(d = 1, 2, 3)$ denoted the degree of a polynomial.

All variables used in the estimation are standardized, since estimated coefficient values should be compared on the same basis, which is common practice in a real data analysis with CCA. Males and females are estimated separately from the result of Fig.2 which show that they may have different preferences in migration routes into the Barents Sea.

The best variables for $\boldsymbol{y}$ and best expression form of $\boldsymbol{w}(z_1, z_2, z_3, z_4)$ were selected by the Bayesian information criterion (BIC) in Schwarz (1978), and the estimation results are shown in Table 2.

Quadratic expressions for latitude / longitude and calendar day and the linear expression for year were selected as the best model in male. Cubic expression for latitude / longitude and year and the liner expression for the calendar day were selected in female.

Canonical correlations are positive, 0.350 and 0.379, in male and female, respectively, which signifies that canonical variates u and v are positively related in both sexes.

From the result in male, BT1 and BT3 have relatively large coefficient values, 0.544 and 0.535, therefore these two variables affect canonical variate u more than length, similarly z_3 and z_4 in year and calendar day, −0.653 and 0.627, respectively, affect v more than other variables, which indicate that males get fatter as fall approaches, whereas their blubber thickness are diminishing year by year over the study period.

Table 2. Estimation result by length, BT1 and BT3.

		Male	Female
Length	y_1	0.154	-0.890
BT1	y_2	0.544	-0.247
BT3	y_3	0.535	0.055
Latitude(z_1)	z_1	0.123	-2.270
Longitude(z_2)	z_2	-0.212	-0.626
	z_1^2	0.049	0.234
	z_2^2	-0.049	-0.090
	$z_1 z_2$	0.218	-0.223
	z_1^3		0.304
	z_2^3		0.118
	$z_1^2 z_2$		0.084
	$z_1 z_2^2$		0.661
Year	z_3	-0.653	0.459
	z_3^2		0.029
	z_3^3		-0.113
Calendar Day	z_4	0.627	0.127
	z_4^2	0.142	
	z_4^3		
Canonical Correlation		0.350	0.379

In the female, the effect of length on u and those of z_1, z_2, $z_1 z_2^2$, and z_3 on v are relatively strong. From coefficient values -0.890 in length, -2.270 in z_1, and -0.626 in z_2, large female whales might migrate in the northeastern part of the study area, however this interpretation is obscure since females are more complicated estimation results where that positive and negative coefficients are mixed in the cubic expression in geography (latitude / longitude). For more clear interpretation of results, estimated varying coefficients are graphically expressed by geography, year, and calendar day in Fig.3 and Fig.4.

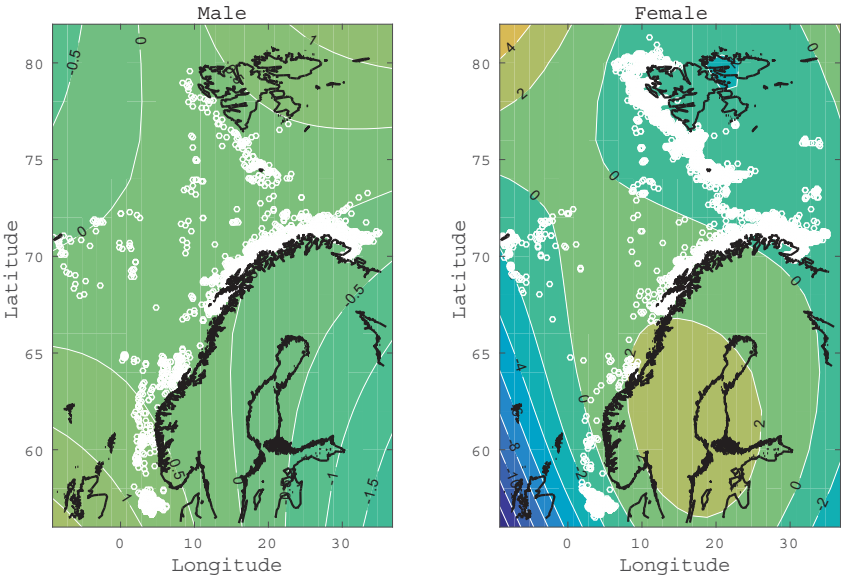

Fig. 3. Varying coefficient cubic plane curves (length, BT1 & BT3).

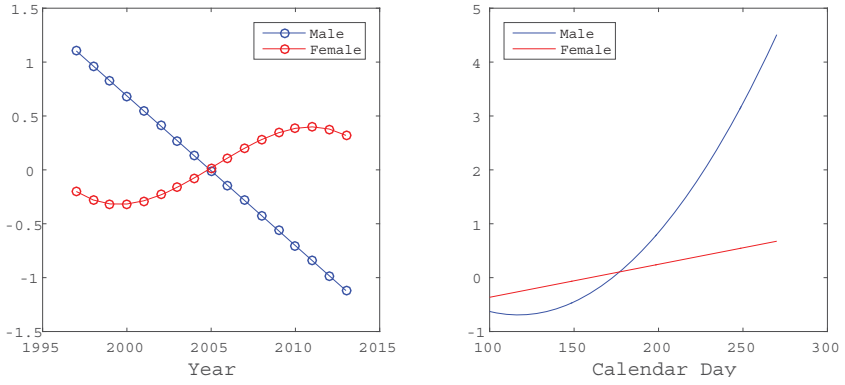

Fig. 4. Varying coefficient plane curves (length, BT1 & BT3).

Fig.3 shows estimated varying coefficients cubic plane curves by sex. White markers are actual catching points (similar to Fig.2), and coefficient values are showed by contour plots which become higher in warmer colored areas.

Contours take values between -1.5 and 1 and are almost flat in male, meaning that body condition considered by length, BT1, and BT3 are not much different in any geographical areas in males. In females, although contours are between -14 and 4 on the map, they are between -2 and 2 at actual observation points, meaning that female body conditions are not much different in observed areas either. In females, the low contour in dark blue at the bottom left in the map, Fig.3 and the high in yellow at the top, might signify habits of whales that they migrate from south with

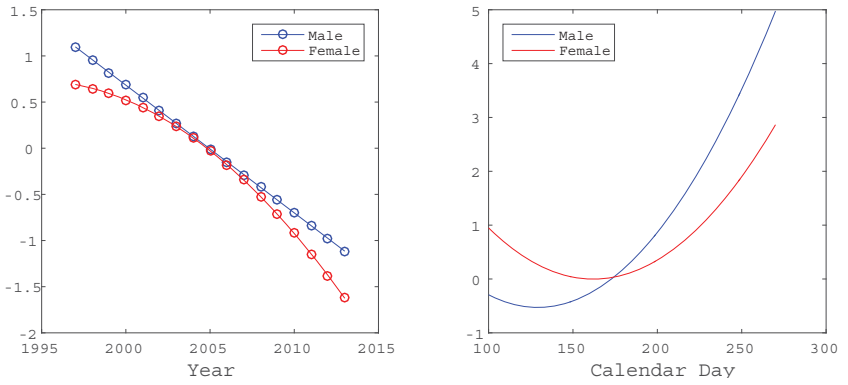

Fig. 5. Varying coefficient plane curves (BT1 & BT3).

hunger and move northward to take enough nourishment in Barents Sea, or might take extreme large or small values since $\boldsymbol{w}_1(z_1, z_2)$ have a cubic expression.

Since liner and cubic expressions for years selected by BIC in male and female, respectively; varying coefficient plane curves in Fig.4 take exactly liner and quadratic forms. Body conditions decrease ever year in male as we already saw the negative effect, -0.653, in Table 2, while the female has a positive slop during 2000-2010 however effects between -0.5 and 0.5 are not large and are negative after 2010. Solvang *et al*. (2016) reports annual year-by-year reduction in the whale's body condition during the study period. In our analyses (Fig.4) the females do not follow the same trend, presumably because the length is included as the whale's body condition along with BT1 and BT3 in our model. Fig.5 shows the year effect on BT1 and BT3 (without the length included), and the effect is negative and similar to the result in Solvang *et al*. (2016). The estimation result about Fig.5 is put in Appendix B.

The calendar day has positive effects on both male and female body conditions in Fig.4, which signifies that whales are nourished and deposit fat reserves in the blubber during summer in the Barents Sea. Fig.5 also shows positive calendar day effects in both male and female which corresponds to the result in Solvang *et al*. (2016).

Briefly summarized, or estimation results indicate that minke whale body condition is affected by geography in females and by chronology in males, however the geographical effect seems not so strong.

4. Conclusion

In this paper, we proposed CCA for geographical and chronological data with multiple response variables by innovating the varying coefficient model. Varying coefficients were estimated by linear model assumed to have linear, quadratic, or cubic expression, and estimation results in the best model chosen by the BIC were presented by contour maps and line plots which made interpretation of the estimation result easy and clear.

From the estimation results of minke whales body condition data, male and female whales gain their body condition as fall approaches, which is the well known as their general habits in the Barents Sea; the nourishment during summer result in fat deposition and leads to fatter body conditions in the fall. Windsland *et al*. (2007) suggested the possibility of food reduction for whales caused by ecological change in Barents Sea, therefore negative year effects in Fig.4 and Fig.5 might arise and express the circumstance of food reduction.

We created two synthesis variables from p-variate vector of response variables and k-variate vector of explanatory variables. As an extension model, more than two synthesis variables can be created, which expands the availability of our model.

Acknowledgements

The author wishes to thank two reviewers for their helpful suggestions. The author's research was supported by JSPS (Japan society for the promotion of science) KAKENHI, Grant-in-Aid for Scientific Research(C), #16K00048, 2016–2019.

Appendix A. Relationship between Our Estimation Method and CCA

Let $\boldsymbol{Y}$ and $\boldsymbol{X}$ be $n \times p$ and $n \times q$ matrices defined by $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and $\boldsymbol{X} = (\boldsymbol{x}_1(z_1), \ldots, \boldsymbol{x}_n(z_n))'$, respectively. Note that $\boldsymbol{Y} - \mathbf{1}_n\bar{\boldsymbol{y}}' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}$ and $\boldsymbol{X} - \mathbf{1}_n\bar{\boldsymbol{x}}' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}$, where $\mathbf{1}_n$ is an n-variate vector of ones and $\boldsymbol{J}_n = n^{-1}\mathbf{1}_n\mathbf{1}_n'$. By using these equations and the properties of $\boldsymbol{I}_n - \boldsymbol{J}_n$ that $(\boldsymbol{I}_n - \boldsymbol{J}_n)' = \boldsymbol{I}_n - \boldsymbol{J}_n$ and $(\boldsymbol{I}_n - \boldsymbol{J}_n)^2 = \boldsymbol{I}_n - \boldsymbol{J}_n$, $\boldsymbol{S}_{yy}$, $\boldsymbol{S}_{xx}$ and $\boldsymbol{S}_{yx}$ in (**??**) can be rewritten as

$$\boldsymbol{S}_{yy} = \frac{1}{n-1}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}, \quad \boldsymbol{S}_{yx} = \frac{1}{n-1}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}, \quad \boldsymbol{S}_{xx} = \frac{1}{n-1}\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}. \tag{A.1}$$

Let $u_i = \boldsymbol{\alpha}'(\boldsymbol{y}_i - \bar{\boldsymbol{y}})$ and $v_i = \boldsymbol{\theta}'\{\boldsymbol{x}_i(z_i) - \bar{\boldsymbol{x}}\}$ $(i = 1, \ldots, n)$. Then, it is easy to see that $\boldsymbol{u} = (u_1, \ldots, u_n)' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}\boldsymbol{\alpha}$ and $\boldsymbol{v} = (v_1, \ldots, v_n)' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}\boldsymbol{\theta}$. Recall that $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are restricted to $\boldsymbol{\alpha}'\boldsymbol{S}_{yy}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\theta}'\boldsymbol{S}_{xx}\boldsymbol{\theta} = 1$, respectively. By uisng these restrictions, the definitions of $\boldsymbol{S}_{yy}$, $\boldsymbol{S}_{yx}$ and $\boldsymbol{S}_{xx}$ in (**??**), and $\boldsymbol{u}$ and $\boldsymbol{v}$, $F(\boldsymbol{\alpha}, \boldsymbol{\theta})$ in (**??**) can be rewritten as

$$\begin{aligned} F(\boldsymbol{\alpha}, \boldsymbol{\theta}) &= (\boldsymbol{u} - \boldsymbol{v})'(\boldsymbol{u} - \boldsymbol{v}) = \{(\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{Y}\boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\theta})\}'\{(\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{Y}\boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\theta})\} \\ &= \boldsymbol{\alpha}'\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}\boldsymbol{\alpha} + \boldsymbol{\theta}'\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X} - 2\boldsymbol{\alpha}'\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}\boldsymbol{\theta} \\ &= (n-1)\left(\boldsymbol{\alpha}'\boldsymbol{S}_{yy}\boldsymbol{\alpha} + \boldsymbol{\theta}'\boldsymbol{S}_{xx}\boldsymbol{\theta} - 2\boldsymbol{\alpha}'\boldsymbol{S}_{yx}\boldsymbol{\theta}\right) = 2(n-1)\left(1 - \boldsymbol{\alpha}'\boldsymbol{S}_{yx}\boldsymbol{\theta}\right). \end{aligned}$$

Hence, the minimization problem in (**??**) can be sovled as

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\alpha}\in\mathcal{A},\ \boldsymbol{\theta}\in\mathcal{T}} \mathrm{RSS}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\alpha}\in\mathcal{A},\ \boldsymbol{\theta}\in\mathcal{T}} \boldsymbol{\alpha}'\boldsymbol{S}_{yx}\boldsymbol{\theta} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^p,\ \boldsymbol{\theta}\in\mathbb{R}^q} \frac{\boldsymbol{\alpha}'\boldsymbol{S}_{yx}\boldsymbol{\theta}}{\sqrt{\boldsymbol{\alpha}'\boldsymbol{S}_{yy}\boldsymbol{\alpha}\ \boldsymbol{\theta}'\boldsymbol{S}_{xx}\boldsymbol{\theta}}}.$$

The above maximization problem is equal to that in CCA.

Appendix B. Estimation result by BT1 and BT3.

The Table B.3. shows the estimation result when $\boldsymbol{y} = (y_1, y_2)' =$ ("BT1", "BT3"). As well as the estimation in Table 2, the best variables for $\boldsymbol{y}$ and best expression form of $\boldsymbol{w}(z_1, z_2, z_3, z_4)$ were selected by BIC. Both BT1 and BT2 are selected in male, while only BT1 is selected in female. Since $\boldsymbol{y}$ does not need to be synthesized in female, the model is just the regression model with one response variable, therefore the coefficient of y_1 takes 1.000 in female. Varying coefficient plane curves in Fig.5 are described by the estimation result in Table B.3.

Table B.3. Estimation result by BT1 and BT3.

		Male	Female
BT1	y_1	0.546	1.000
BT3	y_2	0.588	
Latitude(z_1)	z_1	-0.126	1.668
Longitude(z_2)	z_2	-0.126	0.205
	z_1^2		0.176
	z_2^2		0.218
	z_1z_2		1.206
	z_1^3		
	z_2^3		
	$z_1^2z_2$		
	$z_1z_2^2$		
Year	z_3	-0.650	-0.694
	z_3^2		-0.164
	z_3^3		
Calendar Day	z_4	0.613	-0.026
	z_4^2	0.178	0.107
	z_4^3		
Canonical Correlation		0.345	0.295

References

1. Yamamura M, Fukui K, Yanagihara H. Illustration of the varying coefficient model for a tree growth analysis from the age and space perspectives. *FORMATH* 2016;**15** (in press).
2. Yoshimoto A, Kamo K, Yanagihara H. *Environmental data analysis by R*. Tokyo: Asakura syoten; 2012 (in Japanese).
3. Hastie T, Tibshirani R. Varying-coefficient models. *J Roy Statist Soc Ser B* 1993;**B55**:757-796.
4. Tonda T, Satoh K, Yanagihara H. Statistical inference on a varying coefficient surface using interaction model for spatial data. *Japanese J Appl Statist* 2010;**39**:59-70 (in Japanese).
5. Hashiyama, Y, Yanagihara H, Fujikoshi Y. Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification. *Linear Algebra Appl* 2014;**455**:82-106.
6. Hotelling H. Relations between two sets of variates. *Biometrika* 1936;**28**:321-377.
7. Leurgans SE, Moyeed RA, Silverman BW. Canonical correlation analysis when the data are curves. *J Roy Statist Soc Ser B* 1993;**55**:725-740.
8. Dubin JA, Müller H. Dynamical correlation for multivariate longitudinal data. *J Amer Statist Assoc* 2005;**100**:872-881.
9. Timm HN. *Applied multivariate analysis*. New York: Springer-Verlag; 2002.
10. Fujikoshi, Y, Sakurai T, Kanda S, Sugiyama T. Bootstrap information criterion for selection of variables in canonical correlation analysis. *J Inst Sci Eng ChuoUniv* 2008; **14**:31-49 (in Japanese).
11. Solvang HK, Yanagihara H, Øien N, Haug T. Temporal and geographical variation in body condition of common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the northeast Atlantic. *TR No 16-05*, *Statistical Research Group*, *Hiroshima University*, Hiroshima; 2016.
12. Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978;**6**:461-464.
13. Windsland K, Lindstrøm U, Nilssen K. T., Haug T. Relative abundance and size composition of prey in the common minke whale diet in selected areas of the northeast Atlantic during 2000-04. *J Cetacean Res Manag* 2008;**9**:167-178.